第33卷 第15期 计 算 机 工 程 2007年8月
Vol.33 No.15 Computer Engineering August 2007
·软件技术与数据库· 文章编号：1000—3428(2007)15—0076—03 文献标识码：A 中图分类号：TP393

科技数据网格中基于事件驱动的异步消息传输模型

吴开超[1]，李加升[2]，肖 云[1]，周园春[1]，阎保平[1]
(1. 中国科学院计算机网络信息中心，北京 100080；2. 湖南益阳职业技术学院，益阳 413000)

摘 要：典型数据网格实例在异构节点之间的数据传输效率低效。为了克服其缺点，提出了一个基于事件驱动的异步消息传输模型，给出了相关定义，论述了该模型的体系结构，并应用于科学数据网格中数据访问中间件当中。初步应用表明，该文提出的基于事件驱动的异步传输模型方案，对于具有数据海量性的科学数据网格具有重要意义。

关键词：科学数据网格；事件驱动；异步消息；数据访问中间件

Event-driven Based Asynchronous Message Pass Model for Scientific Data Grid

WU Kai-chao[1], LI Jia-sheng[2], XIAO Yun[1], ZHOU Yuan-chun[1], YAN Bao-ping[1]
(1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100080;
2. Yiyang Vocational & Technical College, Yiyang 413000)

【Abstract】The data exchange between heterogeneous nodes in the representative instances of data grid has low efficiency. In order to overcome the disadvantage of the data grid, an event-driven based asynchronous message pass model is proposed. The related definitions and the architecture of this model are discussed. The data access middleware of SDG based on this model is also developed. The initial usage shows the event-driven based synchronous message pass model is important for SDG, which has mass data.

【Key words】scientific data grid; event-driven; asynchronous message; data access middleware

1 概述

科学数据网格(scientific data grid, SDG)[1]是在中国科学院科学数据库项目中 20 年来积累的海量科学数据资源的基础上，以促进数据共享与开发利用为目的，以先进的数据网格[2]技术为手段，连接分布在全国的 40 多个研究所而建立的一个分布海量数据的一体化网格数据访问、存储、传输、管理与服务架构和环境。它以科学数据管理为中心，面向底层屏蔽网络中各种异构存储和数据资源，面向上层应用提供易于使用的统一访问接口，建立虚拟组织内部数据的统一共享和管理，为用户提供一体化的数据管理和高性能处理服务。

目前国际上主要的数据网格框架[3,4]都是以Globus[5]为基础，根据不同的引用实例来扩展相应的中间件，这样通用性高，也相对比较稳定。所以在具体研究和开发科学数据网格时，也采用了Globus作为底层系统开发平台。但是由于科学数据网格中的海量数据分布于地理上分布的节点，并为地理上分散的网格用户所共享，所以要访问这些海量的数据资源，必须要有一种有效的通信手段，保证不同节点的数据资源的传输、共享的有效性。由于Globus中数据访问模块主要解决计算网格环境下的资源管理，对上层应用开发支持明显不够，对大数据量的传输效率也低，从而影响响应速度，而且缺少对细粒度数据项的管理功能。另外虽然GridFTP[6,7] 提供了一个到各种存储系统的统一的接口，使得可对任何存储系统进行访问；但是GridFTP在访问及复制数据和传输时该体系结构的客户机和待定的存储系统客户机库和协议之间的转换以及不同的存储系统之间的数据传输时格式的转化使得性能严重下降；对于设计者来说建立这样一个接口支持大量不同的存储系统是非常复杂的。

本文根据科学数据网格的特点，对科学数据网格中数据传输的需求进行了分析，在科学数据网格结构基于 Globus 的前提下提出了一种基于事件驱动的异步消息传输机制，并应用于该网格中的数据访问中间件中。

2 相关定义和概念

定义 1 (事件) 事件是代表另一对象变化的抽象对象。

定义 2 (事件驱动) 事件驱动是指根据事件的状态决定执行流程的程序运行模式。它可以表示为一个二元组(S, R)，其中，S 是事件状态集合；R 是运行模式集合。

定义 3 (消息) 消息是对传递数据对象的描述，可以用二元组来表示M:=(H,B)，其中，H表示消息头；B表示消息体。其具体的样式用Abstract Syntax Notation One(ASN.1)[8]来描述，格式如下：

```
message format :=SQEQUENCE{
Version,/*消息版本*/
Type,/*消息类型*/
ID,/*消息 ID，通过随机数产生*/
GID,/*组消息 ID*/
Transport,/*传输方式，如 Socket、SSL 等*/
```

基金项目：国家“863”计划基金资助项目(2002AA104240)；中科院“十五”信息化建设基金资助重大项目(INF105-SDB)

作者简介：吴开超(1970－)，男，博士研究生，主研方向：数据网格；李加升，副教授；肖 云，研究员；周园春，博士研究生；阎保平，研究员、博士生导师

收稿日期：2006-09-01 **E-mail：**yczhou@sdb.cnic.cn

Source,/*消息来源*/

Destination,/*消息目的地址*/

Length,/*消息体的长度，以 4 字节整数表示*/

Message_body/*消息体*/

}

一旦消息格式中指明了消息传输的方式，消息接受器就可以正确的获取消息的内容。

定义 4 (处理器) 处理器是指完成具体事件的处理代码。

定义 5 (任务) 任务是模型的基本工作单位。它是一个包括要完成工作的描述以及完成该工作所需要的数据的消息体。一个任务 T 可以用二元组来表达，T:=(M,P)，其中，M 是任务 P 中消息的结合；O 是任务 P 中处理器的集合。

3 基于事件驱动的异步消息传输模型

3.1 模型介绍

基于事件驱动的异步消息传模型机制(见图 1)是一种轻型的框架，它是基于任务/处理器的事件驱动模型，采用了标准事件驱动应用框架的扩展(事件驱动应用框架+异步消息传输)方式，并提供了简洁有效的事件处理接口。该模型是按层次组织的，自下而上可以分为 4 层：基础层(fabric layer)，传输层(transport layer)，核心层(core layer)，应用层(application layer)也即处理器层。

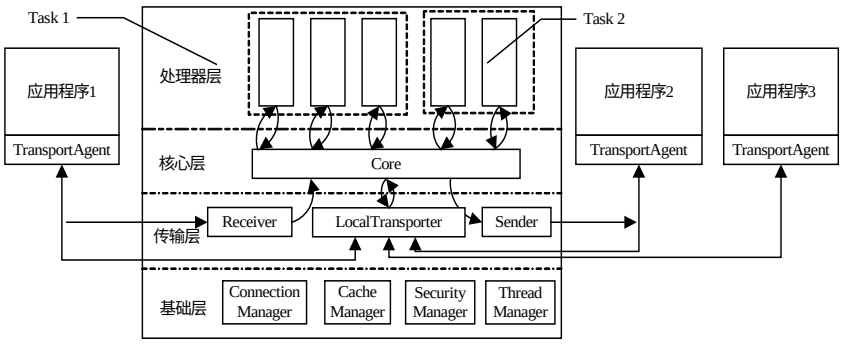

图 1 异步消息传输模型的体系结构

基础层用来完成模型的基础功能，包括安全管理(security management)、连接管理(connection management)、cache 管理(cache management)、线程管理(thread management)等，目前主要借助于 globus 的基础功能。

传输层的主要功能是实现不同节点之间的通信，它可支持多种传输方式，比如普通的 socket 传输、SSL 传输、SOAP 消息的传输等。

核心层的功能是完成应用层处理器之间的事件和消息的分布，在分布的过程为了得到最优的结果，需要依靠科学数据网格中的信息服务中间件提供的元目录信息，并根据各个站点的状况把消息分布到需要的站点上。

应用层主要是一些处理器组成，它的主要任务是通过相应的处理器完成实际的处理功能，这些实际的功能可以是模型本身所拥有的系统任务，也可以是应用开发人员自己开发的任务部署到应用层的，所以从这个含义上来说，应用层又可以看作是处理器的容器。

传输模型中的消息是通过 Sender/Receiver 来发送和接收，Sender/Receiver 在各个站点的传输是异步的，也是对等的，即各个站点的样式是 P2P 模式。另外为了提高处理的速度，对于本地的任务模型专门提供了 LocalTransporter 的传输接口。

由于该模型采用的是异步消息传输，因此对消息的处理是模型的核心功能。图 2 给出的是消息处理框架结构。

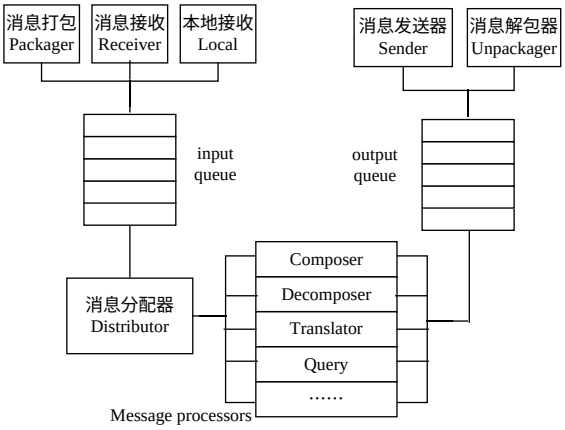

图 2 消息处理流程

其主要模块处理的算法如下：

PackMessage(msg){/*对消息进行打包，目的是把消息的局部 ID 转换成全局的 ID*/

取出 msg 的 ID;

If (msg.ID is LocalID){

把 msg 的局部 ID 转化成全局的 ID;

将全局的 ID 存入到 msg.ID 中;

}

return msg;

}

ReceiveMessage(msg) {/*站点收到消息时，该操作被触发*/

newmsg=PackMessage(msg);

将 newmsg 放入到输入队列 InQueue;

}

ProcessMessage(){/*消息处理线程，可与主线程并行执行，处理消息输入队列 InQueue 中的通知消息，并根据消息的类型把消息分到相应的消息处理器中，消息处理器处理完消息之后把相应的结果消息放到消息输出队列 OutQueue 中*/

while(true){

while(InQueue is not empty){

swith(msg.TYPE){

case GROUP_MESSAGE:

Deomposer(msg);/*根据元目录信息把消息分成若干个子消息，以便送到相应的站点去执行*/

把分解后的消息送到消息输出队列 OutQueue 中;

case REQUEST_MESSAGE:

Dispatrch(msg.ID,msg);/*把消息分配给消息 ID 属于的相应任务处理器，如查询处理器*/

将处理后的结果消息送到消息输出队列 OutQueue 中;

case RESULT_MESSAG:

Composer(msg);/*组装各个发送过来的结果消息*/

把组装后的结果消息送到消息输出队列 OutQueue 中;

}

把处理过的 msg 放到输出队列 OutQueue 中;

}

}

}

```
UnpackMessage(msg){/*对消息进行解包，目的是把消息的全局的 ID 转换成局部 ID*/
取出 msg 的 ID;
If (msg.ID is GlobalID){
把 msg 的全局 ID 转化成局部的 ID;
将局部的 ID 存入到 msg.ID 中;
 }
return msg;
}

SendMessge(){/*根据消息的目的地址发送该消息*/
  while (OutQueue is not empty){
从 OutQueue 中取出消息 msg;
newmsg=UnpackMessage(msg);
Send(newmsg);
  }
}
```

3.2 模型分析

试验采用了 5 台 Linux 服务器，比较 GridFtp 和本模型的传输效率，具体结果如表 1。

表 1　本模型和 GridFtp 的传输性能比较

模型	文件大小/MB	传输时间/s
GridFtp (TCP Buffer size=8KB)	40	4.523
	80	10.105
	120	31.256
本文提出的模型	40	3.863
	80	9.586
	120	31.213

从表中可以看出，本文提出的模型在小文件传输上具有优势，原因在于：

(1)在传输方式上采用异步方式，发送者向接收者发送消息后继续进行后面的动作，不管接收者是否接收到该消息，提高了消息传输的并发度；

(2)消息传输和消息的分配一体性，在基于事件驱动的异步消息传输模型中，把消息的传输和消息的分配结合起来，当消息接收器收到消息后能把消息分配到适当的消息处理器中；

(3)在传输过程中封包和解包的过程少。

本文提出的模型比较适合与数据网格的其他服务相结合使用，为上层应用开发提供了很好的支持，比如该模型在科学数据网格分布式查询处理方面和网格资源监控方面使用更加方便。如果是单纯的大文件传输优势不大。

4 应用实例

科学数据网格是中国网格的一个部分，其软件模块主要有两大部分组成：应用软件和网格中间件。应用软件是为了更好地管理各个网格站点系统，它的实现主要依托中间件提供的一些服务和接口。中间件模块包括 4 大部分：

(1)信息服务中间件，它依托 MDS/LDAP，存储着数据服务所需要的核心目录信息，为数据访问中间件提供目录服务；

(2)安全服务中间件，它为系统的安全访问数据提供认证/授权信息，为安全的数据访问提供保证；

(3)存储服务中间件，它主要是为用户提供一个访问多种异构存储系统的统一接口，它屏蔽了存储系统的异构特性，支持广域网络环境下多种数据源的访问；

(4)数据访问中间件，是建立在安全体系构建的安全环境下，对外提供统一的数据访问接口，屏蔽分布式环境下数据资源的多样性和异构性，以去除数据孤岛/信息孤岛，实现数据资源的共享和集成。

在科学数据网格中涉及数据访问和传输的主要是数据访问中间件，所以在开发数据访问中间件时利用基于事件驱动的异步消息传输模型，并且基于该中间件开发了一些基本的应用，如数据的查询和集成以及服务的监控，取得了良好的效果。

5 结束语

对海量数据访问有效性的研究是数据网格研究的热点之一。本文为科学数据网格中的数据通信提出了一种基于事件驱动的异步消息传输模型，它能够很好地支持科学网格这种开放性、分布性、协作型和动态性的协同环境，并为之提供了一种清晰的方法用于连接应用中不同的组件；另外该模型有助于把一个需要长时间运行的任务分解成多个事务，从而增加其有效性，能为上层应用开发提供很好的支持；该模型采用异步的消息传递方式，提高了处理速度和系统的可靠性。进一步研究的方向主要是考虑模型中事件的优先级和任务的可移动性等。

参考文献

1 Nan Kai, Yan Baoping. Introduction to Scientific Data Grid[C]//Proc. of the APAN Grid Workshop. 2002.

2 Chervenak A, Foster I, Kesselman C, et al. The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets[J]. Journal of Network and Computer Applications, 2001, 23(3): 187-200.

3 武秀川, 胡　亮, 鞠九滨. 数据网格的数据管理策略[J]. 小型微型计算机系统, 2004, 25(1): 98-102.

4 Stockinger H. Distributed Database Management Systems and the Data Grid[C]//Proc. of the 18th IEEE Symposium on Mass Storage Systems and 9th NASA Goddard Conference on Mass Storage Systems and Technologies, San Diego. 2001: 17-20.

5 Foster I, Kesselman C. Globus: A Metacomputing Infrastructure Toolkit[J]. International Journal of Supercomputer Applications, 1997, 11(2): 115-128.

6 Allcock W. GridFTP Protocol Specification (Global Grid Forum Recommendation GFD.20)[Z]. (2003). http://www.globus.org/ research /papers/GFD-R.0201.pdf.

7 Allcock B, Bester J, Bresnahan J, et al. Data Management and Transfer in High Performance Computational Grid Environments[J]. Parallel Computing Journal, 2002, 28(5): 749-771.

8 Larmouth J. ASN.1 Complete[Z]. (2003). http://www.oss.com/asn.1/ larmouth.html.